



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Discourse for Machine Translation

Citation for published version:

Webber, BL 2014, Discourse for Machine Translation. in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 28, Cape Panwa Hotel, Phuket, Thailand, December 12-14, 2014*. pp. 27. <<http://aclweb.org/anthology/Y/Y14/Y14-1004.pdf>>

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 28, Cape Panwa Hotel, Phuket, Thailand, December 12-14, 2014

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Discourse for Machine Translation

Bonnie Webber

School of Informatics, University of Edinburgh

Abstract

Statistical Machine Translation is a modern success: Given a source language sentence, SMT finds the most probable target language sentence, based on (1) properties of the source; (2) probabilistic source--target mappings at the level of words, phrases and/or sub-structures; and (3) properties of the target language.

SMT translates individual sentences because the search space even for a single sentence can be vast. But sentences are parts of texts, and texts have properties beyond those of their individual sentences, including:

- document-wide properties, such as style, register, reading level and genre, that are visible in the frequency and distribution of words, word senses, referential forms and syntactic structures;
- patterns of topical or functional sub-structures that mean that frequencies and distributions of words, word senses, referential forms and syntactic structures will vary across a text;
- relations between clauses or between referring expressions that can be signaled explicitly or implicitly, that reflect a text's coherence;
- frequent appeal to reduced expressions that rely on context to
- efficiently convey their message.

Recognizing and deploying these properties promises to improve both fluency and accuracy in SMT -- i.e., whether the sequence of sentences in the target text conveys the same information as those in its source, in as readable a manner. This presentation describes how researchers are attempting to do this, without bringing translation to a halt.